Verification of Single-Peptide Protein Identifications by the Application of Complementary Database Search Algorithms

James G. Rohrbough[1, 2], Linda Breci[3], Nirav Merchant[4], Susan Miller[4], Paul A. Haynes[1, 5]

[1] Department of Biochemistry and Molecular Biophysics, The University of Arizona, Tucson, AZ, USA

[2] Air Force Institute of Technology, Civilian Institutions Programs, Wright-Patterson Air Force Base, OH, USA

[3] Department of Chemistry, The University of Arizona, Tucson, AZ, USA

[4] Arizona Research Laboratories, Biotechnology Computing Facility, The University of Arizona, Tucson, AZ, USA

[5] Bio5 Institute for Collaborative Bioresearch, The University of Arizona, Tucson, AZ, USA

Running Title: Single-Peptide Protein Identification Verification by Dual Search Algorithms

Correspondence: Associate Professor Paul A. Haynes, Department of Biochemistry and Molecular Biophysics, the University of Arizona, Biological Sciences West 349, 1041 East Lowell Avenue, Tucson, Arizona, 85721, USA

E-mail: phaynes@email.arizona.edu

Fax: 1-520-626-4824

Abbreviations: MudPIT, multidimensional protein identification technique; SCX, strong cation exchange

Keywords: Database search algorithms / Protein identification / SEQUEST criteria / Tandem mass spectrometry / XTandem

Abstract

1 Introduction

Protein identifications from complex biological mixtures often involve the application of tandem mass spectrometry techniques. One such technique, known as the Multi-Dimensional Protein Identification Technique, or MudPIT, involves the use of computer search algorithms that automate the process of identifying proteins present in the sample mixture based on mass spectrometry analysis. This technique involves digestion of the protein mixture with a protease such as trypsin, followed by liquid chromatography separation using first a strong cation exchange column followed by a reverse-phase separation. Peptides eluting from these separations are subjected to ionization and fragmentation in the mass spectrometer. The database search algorithms are then used to match the acquired spectra to peptide sequences from a protein database. These algorithms, while helpful, are far from perfect when it comes to accuracy of peptide identifications. These programs identify peptides by comparing the collected spectra to predicted spectra from the database sequences and applying a score to that identification. The peptide with the highest score is the one selected as the identification. The user is able to select a cutoff score or scores above which identifications are kept, and below which identifications are disregarded. When a protein is identified from several unique peptide spectra, the inherent redundancy of identification provides a significant confidence of protein identification, even if the confidence of some of the peptide identifications is low. As the number of peptides assigned to each protein sequence decreases, the confidence of protein identification drops, until we reach the proteins identified from one unique peptide sequence. These proteins rely completely on the ability of the database search algorithm, and the applied score cutoff parameters for identification. We propose a system of analysis that utilizes the consensus between two popular search algorithms, SEQUEST and XTandem, to increase the confidence of protein identifications from single peptides, while minimizing false-positive identifications.

There are many example in current literature of proteomic analyses performed by application of the MudPIT technique [1-5]. However, there is no consensus on the search parameters used for the database search algorithm, or the treatment of the proteins identified from single peptides. It is not correct to simply disregard single-peptide matches [3] because such peptides may be the only detectable peptide from an enzymatic digest, and therefore perfectly valid for identification purposes. It is equally incorrect to include all proteins identified from single peptides because of the variability in protein identification from poor mass spectra, resulting in a high rate of false-positive identifications.

There have been numerous attempts to validate protein identifications from current database search algorithms. Many of these involve statistical modeling, such as the linear discriminate analysis used to determine the accuracy of search algorithm assignments [6], or the Qscore algorithm using a probabilistic scoring system and analysis of false-positive identification rates using a reverse database [7]. Some of the validation schemes utilize manipulation of search parameters to achieve higher confidence of protein identifications [8; 9], as well as utilization of the tryptic status of peptides as an additional level of validation [10-13]. Yet another approach involves the application of a machine learning algorithm, known as the support vector machine (SVM), that uses mixtures of known proteins to train the SVM to distinguish between correct and incorrect peptide identifications by SEQUEST [14]. Some approaches use the inclusion of orthogonal parameters such as exact mass measurements of selected peptides [15], although this requires the use of a mass spectrometer capable of such exact measurements. In addition, some have used liquid chromatography information to match peptide elution times to predicted sequences [16].

There are published reports involving proteomic analysis in which the final results are in the form of a consensus between the output from two different search algorithms [17]. This study relied on the use of SEQUEST and Mascot, both of which are commercial products which require purchase of a license. However, neither this report, nor any of those mentioned above, specifically address the issue of improving the confidence rate of assignment for proteins identified from a single peptide.

Our aim in this study was to develop a set of software tools that would enable us to achieve much higher confidence in our single peptide based protein identifications. Our specific goal was to reach 95% confidence of assignment, or greater, for both single and multiple peptide based protein identifications, using only freely available, open-source software in addition to our existing SEQUEST analysis platform. As a consequence, all software tools developed and used in this project are made freely available via our lab website.

Data were acquired from MudPIT analyses of yeast (*S. cerevisiae*) mixed organelle lysate and rice (*O. sativa*) tissue samples. These were used to optimize a set of SEQUEST cutoff parameters which give a greater than 95% confidence that the assigned proteins from multiple peptide matches are valid, assessed by using reversed database searching [7]. The spectra corresponding to the single peptide matches from the initial SEQUEST search are then sorted and reanalyzed by a complementary search with the XTandem algorithm. Single peptide identifications that are matched by both search algorithms are accepted as valid as we demonstrate that they also have at least a 95%

confidence level. The format of the final result output is an Excel spreadsheet indicating the consensus of both the DTAselect filtered sequest results and the reanalysis of the single peptide matches using XTandem, with revised summary totals calculated. We show that this procedure is both reproducible across replicate analyses of the same sample, and equally applicable to samples from distinctly different biological starting materials.

2 Materials and Methods

2.1 Preparation of yeast mixed organelle lysate

The yeast cell samples were processed as described [5]. Briefly, yeast mixed organelle lysate was reduced with dithiothreitol and carbamidomethylated with iodoacetamide. Sample was then digested with endoproteinase Lys-C and trypsinized with Poroszyme immobilized trypsin beads (Applied Biosystems, Framingham, MA, USA). The tryptic digest solution was desalted and purified on a Spec PT C18 solid phase extraction pipette tip (Varian, Lake Forest, CA, USA), dried under vacuum and reconstituted in 0.5% HPLC grade formic acid (Merck, Darmstadt, Germany).

2.2 Preparation of rice leaf and root tissue lysates

Rice plants (*Oryza sativa*, cv. Nipponbare), were grown from seed in a temperature controlled greenhouse under a 12h light 29°C / 12h dark 21°C regime. Humidity was maintained at 30%, and plants were grown in pots containing 50% Sunshine Soil Mix and 50% nitrohumus. Leaf and root samples were collected 50 days after germination and were pooled from multiple plants. Harvested leaves and roots were ground to a fine powder using liquid nitrogen in a mortar and pestle. Protein extracts were prepared using TCA/acetone precipitation, and protease digests of extracted protein were prepared as previously described [18]. Briefly, proteins were denatured in 8M Urea and then sequentially digested by endoproteinase Lys-C and trypsin (Sigma, St. Louis, MO, USA). The resulting digest solution was desalted and purified using C-18 solid phase extraction as described above.

2.2 Nanoflow two-dimensional liquid chromatography- - tandem mass spectrometry (MudPIT)

Analysis of both yeast and rice samples were accomplished by nanoflow two-dimensional liquid chromatography – tandem mass spectrometry, commonly referred to as MudPIT [10], by a previously described method [5]. Briefly, a microbore HPLC system was modified to operate at capillary flow rates using fused silica columns packed with 5μm Zorbax Eclipse XDB C-18 resin (Agilent Technologies, Palo Alto, CA, USA) and 5μm polysulfoethyl-A strong cation exchange resin (PolyLC Inc., Columbia, MD, USA). Samples were introduced onto the column using a Surveyor autosampler. The HPLC column eluted directly into the ESI source of a ThermoFinnigan LCQ-Deca XP Plus ion trap mass spectrometer (Thermo Electron, San Jose, CA, USA). Peptides were eluted in a NH_4HCO_3 gradient at a flow rate of 400 nL/min. A ten salt-step fractionation was performed, for a total of 13 fractions that were generated and analyzed.

2.3 Database searching and false positive rate determination

The entire set of tandem mass spectra collected from all chromatographic steps are searched against an appropriate protein sequence database using SEQUEST (BioWorks version 3.1) (Thermo Electron) [8; 19] and single-peptide matches confirmed by searching the same database using XTandem version 2004.11.15.3 (an open source software, available from the Manitoba Centre for Proteomics at http://www.proteome.ca/opensource.html) [20; 21]. False-positive protein identification rates were calculated from searching against a reversed protein sequence database [7]. The reverse database was produced using an in-house developed perl script.

MS/MS spectra were searched against a database of rice protein sequences (36318 sequences) downloaded from publicly available resources at NCBI (www.ncbi.nlm.nih.gov), and the yeast genome from the *Saccharomyces* Genome Database (www.yeastgenome.org), which was combined with bovine (*Bos taurus)* and equine (*Equus caballus*) genomes for a total of 12976 sequences. Both the rice and yeast databases were supplemented with an in-house contaminants file including trypsin, Lys-C, keratin, albumin, casein and other common laboratory contaminants. SEQUEST search results were filtered using DTA-select (available at http://fields.scripps.edu/DTASelect/) [22] using the indicated cutoff parameters.

2.4 Data manipulation tools

Manipulation of mass spectrometry data was assisted by the use of several perl script programs designed in-house. These scripts include the sub_append.pl script, which combines all SEQUEST-produced .dta files contained in a sub directory into one .dta file. Next is the append.pl script, which is similar to the sub_append script, but instead combines all .dta files in a parent directory into one .dta file. Using these two scripts in sequence produces a single .dta file that contains all of the .dta files from a complete MudPIT run, allowing the complete dataset to be searched using the XTandem program.

To extract those dta files corresponding to SEQUEST single-peptide identifications, the DTA_sorter.pl script is used. This script uses the DTASelect-filter.txt output file and separates all .dta files from a MudPIT run into three folders. The first folder contains all .dta files that correspond to single-peptide identifications (singlexcel). Second is the folder containing all of the remaining unidentified .dta files that correspond to protein identifications in the SEQUEST analysis (inexcel). Last is the folder containing all of the remaining .dta files (notinexcel). To remove .dta files for XTandem searching, one would use the append.pl script on the singlexcel folder, producing a single appended .dta file.

2.5 Algorithm consensus determination

For data comparison purposes, the CommonSingles.pl script compares a standard DTASelect output file (DTASelect-filter.txt) to an XTandem excel table output (obtained using the Global Proteome Machine xml input page at: http://h451.thegpm.org/tandem/thegpm_upview.html). The common singles script produces a modified DTASelect output file that includes all of the single peptides found by XTandem that are also found by SEQUEST. For determining false positive rates, preparation of reverse databases was done using the reverse.pl script. All perl scripts along with usage instructions are available for download at http://proteomics.arl.arizona.edu/perl.html.

3 Results and discussion

3.1 Variability in protein identifications using published SEQUEST parameters

The first step in our analysis was to optimize SEQUEST cutoff parameters to produce a greater than 95% confidence in assignment of multiple peptide based protein identifications. We began this process with a literature survey. There are many different sets of published SEQUEST parameters in the current literature. Figure 1 illustrates the variability in protein identifications based on the analysis of a single yeast MudPIT dataset using five different sets of SEQUEST cutoff scores chosen from published studies. It is apparent that the relatively good agreement between the protein totals identified from multiple peptides indicates that the majority of the variability in the total numbers of protein identifications comes from the single-peptide identifications. The SEQUEST cutoffs parameters used in the references referred to in Figure 1 are listed in Table 1. It is important to note that the data shown for Reference 3 is modified slightly from the published parameters. In this reference, the authors did not include any single-peptide protein identifications. We have included them in the interest of completeness. The cutoff scores in Reference 1 are used in our laboratory as our standard SEQUEST cutoff scores, developed over many years of experience with a very wide range of sample types.

3.2 False positive rates from different published SEQUEST search parameters

Six different sets of MudPIT analysis data were acquired; three replicates of aliquots of a yeast mixed organelle lysate, and three different rice tissue samples, prepared from leaf, root and seed. All six data sets were searched using SEQUEST against both a forward and reversed database, to allow assessment of false positive rates of assignment [7]. All twelve results (six forward and six reversed) were then filtered using each of the five SEQUEST parameter cutoff sets as listed in Table 1.

Figure 2 shows the false positive rates produced by each of the five SEQUEST cutoff scores when applied to the analysis of yeast dataset 1. Table 2 shows the false positive rates produced by the analysis of all six MudPIT datasets using the SEQUEST cutoff parameters from reference 1. In all cases, the largest contributor to the overall false-positive rate is the proteins identified from single-peptides. Reference 1 cutoff scores, which are already in use in our laboratory, produce a multiple-peptide false positive rate below the 5% threshold we are aiming for. None of the other cutoff parameters has all multiple-peptide identifications under a 5% false positive rate. Since the SEQUEST cutoff scores in use in our laboratory reached our goal of 95% confidence in multiple peptide identifications, we decided to use a second database search algorithm specifically for reanalysis of single peptide-based identifications from SEQUEST.

3.3 Development of software tools to sort single-peptide identification spectra

Our plan of validating single-peptide protein identifications using a complementary database search algorithm required the use of an algorithm that was freely available. XTandem provided the desired open-source search algorithm that was easily configured, and performed database searches much faster than SEQUEST. In order to utilize this secondary search program, however, we had to design some perl script-based software tools to assist us. The first program is the DTA_sorter.pl script, which parses out of a larger dataset only those spectra that SEQUEST matched as single-peptide protein identifications. Once the relevant spectra are sorted, they are concatenated by another script into a single .dta spectrum file for use by the XTandem program. These software tools allow us to sort thousands of MudPIT spectra quickly and easily and are available for free download at our website (http://proteomics.arl.arizona.edu).

3.4 Complementary analysis of single-peptide spectra using the XTandem search algorithm

Since we were planning to use XTandem as a second search algorithm, we re-analyzed the complete yeast MudPIT dataset 1 using XTandem to determine a stringency level of result filtering that produced similar output to the SEQUEST data. The main parameter used for results filtering in XTandem is the expectation value (e-value) as determined by the algorithm. As shown in Figure 3, an e-value cutoff of 0.02 produces results that are very similar to those produced using our standard SEQUEST cutoff scores (Ref 1). We also analyzed the false positive rates produced by filtering the XTandem search results for yeast MudPIT dataset 1 at an e-value cutoff of 0.02. The overall false positive rate was 15.6%, which consisted of a 28.6% false positive rate for single peptides and a 4.1% false positive assignment rate for multiple peptide protein identifications (data not shown). This led us to select an XTandem e-value cutoff of 0.02 for use in all further analyses.

3.5 Software development to use SEQUEST and XTandem results for validation of single-peptide protein identifications

Another perl script was developed to automate the comparison of SEQUEST and XTandem results for single-peptides. This program (CommonSingles.pl) produces a final output spreadsheet very similar in format to the DTASelect-filter.txt file produced by the DTASelect program. The output of this script includes revised protein totals, as well as more detailed information regarding which proteins were matched, how many were validated and which were rejected. This script is also available for download on our website.

3.6 Revised results of MudPIT analysis using SEQUEST and XTandem consensus for single-peptide protein identifications

Table 2 shows the false positive rates obtained by using our selected SEQUEST cutoff scores (Ref 1) further sorted by the number of peptides found per protein Table 2 lists the revised numbers of proteins identified using the complementary search algorithm technique we have described here. Each of the yeast samples is from the same yeast culture, with each sample processed identically but separately from the others. Within the yeast samples, there is a high level of reproducibility in the results. When compared to samples prepared from rice tissues, there is a clear difference in false positive rates. However, for all six datasets, we still see the same 95% or greater confidence in multiple-peptide protein identifications. Using the dual algorithm approach outlined above for validation of single peptides identifications, we also see a drastic reduction in the overall false positive rates, and a false positive rate of 0% for single-peptide protein identifications. The resulting percentages of protein identifications retained as a result of verification of single peptides is listed in Table 3. The reanalysis of the yeast MudPIT datasets results in the retention of approximately 80% of all proteins identified by SEQUEST which includes approximately 60% percent of the single-peptide identifications. In contrast, the rice MudPIT datasets show only about 60-70% of the total proteins are retained, which includes 45-50% of the single peptide identifications.

4 Concluding remarks

We have presented a method for verifying proteins identified from a single unique peptide during MudPIT analysis of a complex biological mixture. By validating single-peptide protein identifications using complementary database search algorithms, we can reduce the overall false-positive rates for protein identifications considerably. For the analysis of yeast MudPIT datasets, we are able to produce a revised results output with an overall false positive assignment rate of less than 1%, which still retains 80% of the proteins initially identified. Similary, for analysis of the rice MudPIT datasets, we are able to retain 60-70% of the proteins initially identified, with a revised overall false positive rate less than 1%. This indicates that application of this technique is highly reproducible for the analysis of similar samples, and likely to yield comparable, but slightly different results for samples prepared from different biological sources.

We have developed a technique that can be employed by anyone utilizing a SEQUEST-based proteomic analysis platform, using the XTandem algorithm as a complementary tool for verification of single-peptide protein identifications. We have achieved this using exclusively open-source software, including several data-manipulation software tools developed in our laboratory, all of which are freely available for download. We make these programs available to other users in the spirit of open-source collaboration. We expect that users will modify them to fit their own needs and the continued development of such tools will be a great benefit to the scientific community at large.

Acknowledgements

The authors would like to thank Tim Radabaugh, Mike Galligan, and Fatimah Hickman for their laboratory assistance and helpful discussions. Additional thanks to the Bio5 Institute for Collaborative Bioresearch for funding. *The views expressed in this paper are those of the author and do not reflect the official policy or position of the Air Force, Department of Defense or the United States Government.*

5 References

[1] Breci, L., Hattrup, E., Keeler, M., Letarte, J., *et al.*, *Proteomics* 2005, *5,* 2018-2028.
[2] Benzinger, A., Muster, N., Koch, H.B., Yates, J.R., 3rd, and Hermeking, H., *Mol Cell Proteomics* 2005.
[3] Durr, E., Yu, J., Krasinska, K.M., Carver, L.A., *et al.*, *Nat Biotechnol* 2004, *22,* 985-992.
[4] Phillips, G.R., Anderson, T.R., Florens, L., Gudas, C., *et al.*, *J Neurosci Res* 2004, *78,* 38-48.
[5] Washburn, M.P., Ulaszek, R., Deciu, C., Schieltz, D.M., and Yates, J.R., 3rd, *Anal Chem* 2002, *74,* 1650-1657.
[6] Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R., *Anal Chem* 2002, *74,* 5383-5392.
[7] Moore, R.E., Young, M.K., and Lee, T.D., *J Am Soc Mass Spectrom* 2002, *13,* 378-386.
[8] Eng, J., McCormack, A., and Yates, J.R., III, *J Am Soc Mass Spectrom* 1994, *5,* 976-989.
[9] Yates, J.R., Eng, J., and McCormack, A., *Anal Chem* 1995, *67,* 3202-3210.
[10] Link, A., Eng, J., Schieltz, D., Carmack, E., *et al.*, *Nat Biotechnol* 1999, *17,* 676-682.
[11] Han, D.K., Eng, J., Zhou, H., and Aebersold, R., *Nat Biotechnol* 2001, *19,* 946-951.
[12] Washburn, M., Wolters, D., and Yates, J.r., *Nat Biotechnol* 2001, *19,* 242-247.
[13] Qian, W.J., Liu, T., Monroe, M.E., Strittmatter, E.F., *et al.*, *J Proteome Res* 2005, *4,* 53-62.
[14] Anderson, D.C., Li, W., Payan, D.G., and Noble, W.S., *J Proteome Res* 2003, *2,* 137-146.
[15] Smith, R.D., Anderson, G.A., Lipton, M.S., Pasa-Tolic, L., *et al.*, *Proteomics* 2002, *2,* 513-523.
[16] Strittmatter, E.F., Kangas, L.J., Petritis, K., Mottaz, H.M., *et al.*, *J Proteome Res* 2004, *3,* 760-769.
[17] Resing, K.A., Meyer-Arendt, K., Mendoza, A.M., Aveline-Wolf, L.D., *et al.*, *Anal Chem* 2004, *76,* 3556-3568.
[18] Koller, A., Washburn, M., Lange, B., Andon, N., *et al.*, *Proc Natl Acad Sci U S A* 2002, *99,* 11969-11974.
[19] Yates, J., Eng, J., McCormack, A., and Schieltz, D., *Anal Chem* 1995, *67,* 1426-1436.
[20] Craig, R., and Beavis, R.C., *Rapid Commun Mass Spectrom* 2003, *17,* 2310-2316.
[21] Craig, R., and Beavis, R.C., *Bioinformatics* 2004, *20,* 1466-1467.
[22] Tabb, D., McDonald, W., and Yates, J.R., *J Proteome Res* 2002, *1,* 21-26.

	Minimum X Correlation			
Reference #	**+1 ion**	**+2 ion**	**+3 ion**	**ΔCn**
1	1.8	2.5	3.5	0.1
2	1.8	2.5	3.8	0.08
3	1.5	2.2	3.3	0.07
4	1.8	2.5	3.5	0.08
5	1.9	2.2	3.75	0.1

Table 1. SEQUEST cutoff scores (as filtered by DTASelect) for different published MudPIT studies.

Datasets	Total Proteins Identified	Revised Total Proteins Identified	False Positive (FP) Rates per Number of Peptides Found					Overall FP Rate	Revised Overall FP Rate
			1	Revised 1	2	3	4+		
Yeast									
1	540	445	49.4	0.0	3.7	1.9	0.0	24.4	0.89
2	605	485	50.3	0.0	2.0	0.0	0.0	25.0	0.41
3	522	403	50.2	0.0	1.2	3.9	0.0	26.6	0.74
Rice									
Seed	221	141	41.9	0.0	3.1	0.0	0.0	29.9	0.71
Root	258	174	28.6	0.0	0.0	0.0	0.0	19.4	0.00
Leaf	247	153	59.2	0.0	2.6	0.0	0.0	40.9	0.65

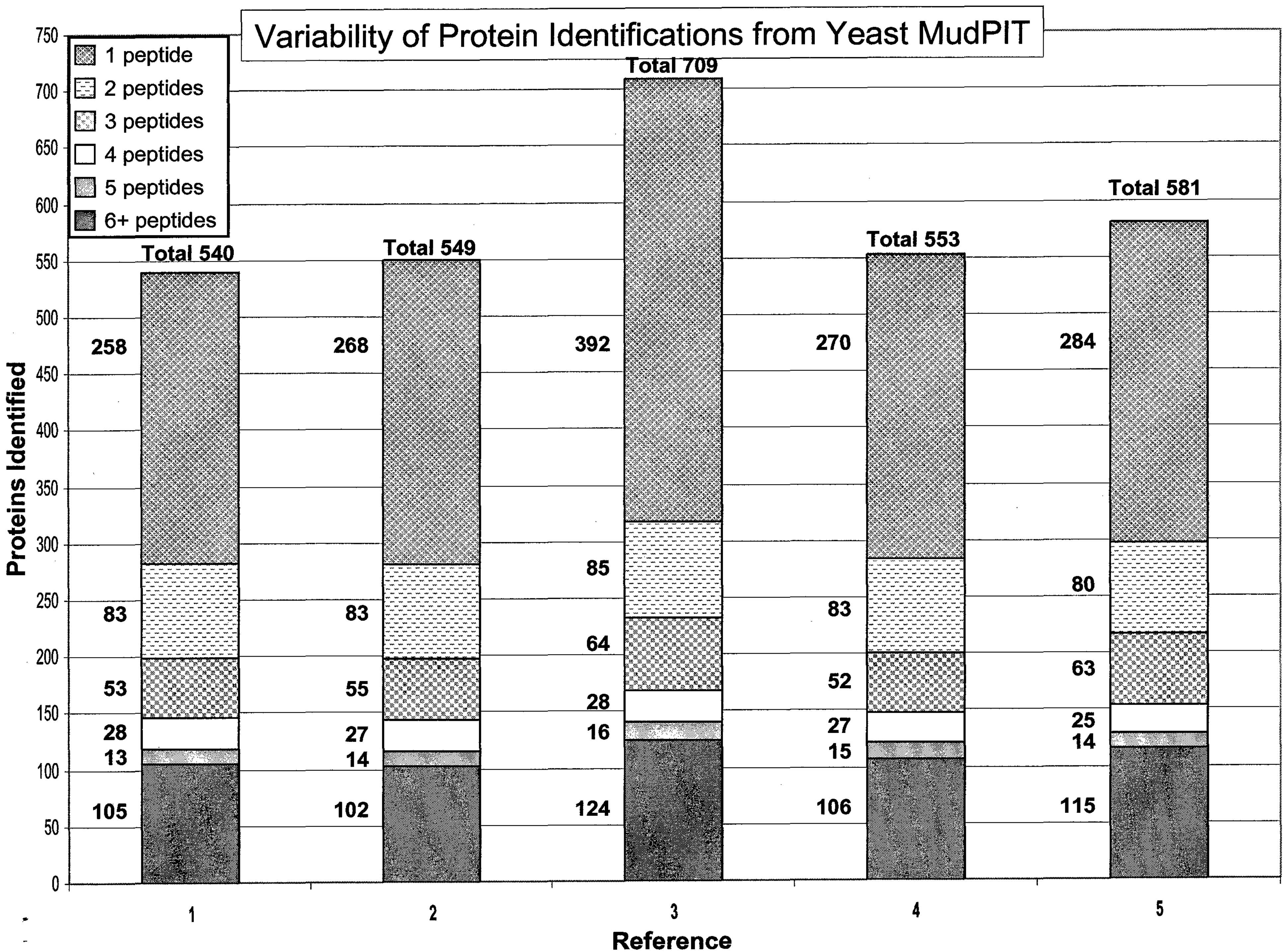
Table 2. False positive rates of Yeast and Rice MudPIT protein identifications using Ref 1 SEQUEST cutoff scores, including revised totals after analysis of single peptide spectra with XTandem (e-value = 0.02). The majority of false positives come from the proteins identified from single peptides. Proteins identified from multiple peptides have a greater than 95% confidence. Revised totals have a greater than 99% confidence.

	% total proteins kept	% singles kept
Yeast		
1	82.4	63.3
2	80.2	59.5
3	77.2	56.1
4	80.1	61.8
Rice		
Seed	63.8	48.4
Root	67.4	52.0
Leaf	61.9	44.4

Table 3. Percent of identified proteins kept using complementary search algorithms.

Figure 1. Variability in protein identifications using different published search criteria. All searches were performed against the same database using the same yeast MudPIT data. Identifications presented here include common contaminates (trypsin, keratin), and are required to be at least partially tryptic. For our purposes, a single peptide is defined as one unique peptide used as the basis for an identification, regardless of the number of times the peptide was identified.

Fig 1



Variability of Protein Identifications from Yeast MudPIT
1 peptide
2 peptides
3 peptides
4 peptides
5 peptides
6+ peptides
Proteins Identified
750
700
650
600
550
500
450
400
350
300
250
200
150
100
50
0
Total 540
258
83
53
28
13
105
Total 549
268
83
55
27
14
102
Total 709
392
85
64
28
16
124
Total 553
270
83
52
27
15
106
Total 581
284
80
63
25
14
115
1
2
3
4
5
Reference

Figure 2. Comparison of false positive protein identification rates using different published SEQUEST search cutoff parameters. Data corresponds to that shown in Figure 1. The total false positive rate includes all identified proteins, regardless of number of unique peptides assigned. The three remaining columns for each reference is the false positive rates for proteins identified from one, two, and three or more unique peptide matches.

Fig 2

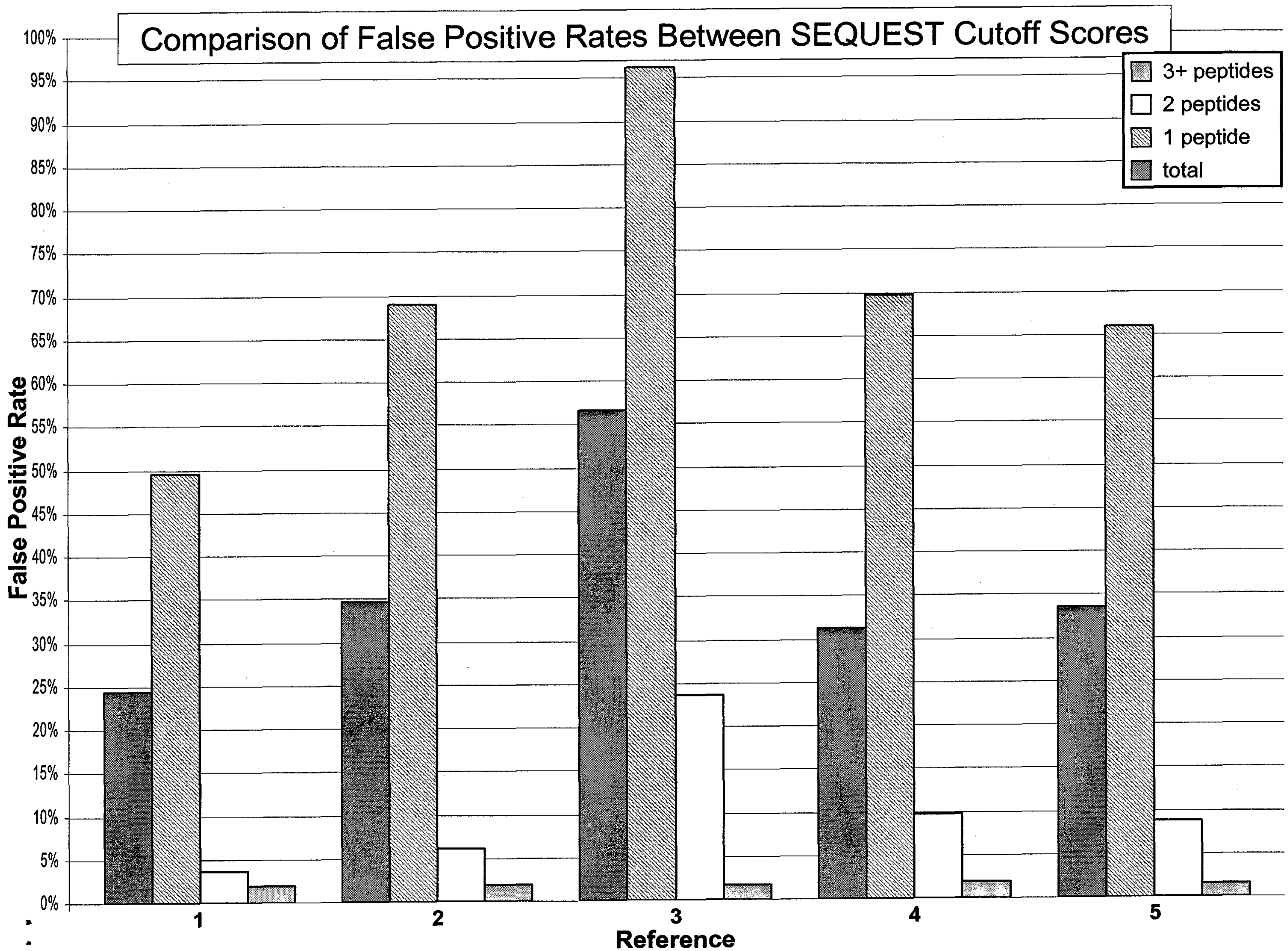
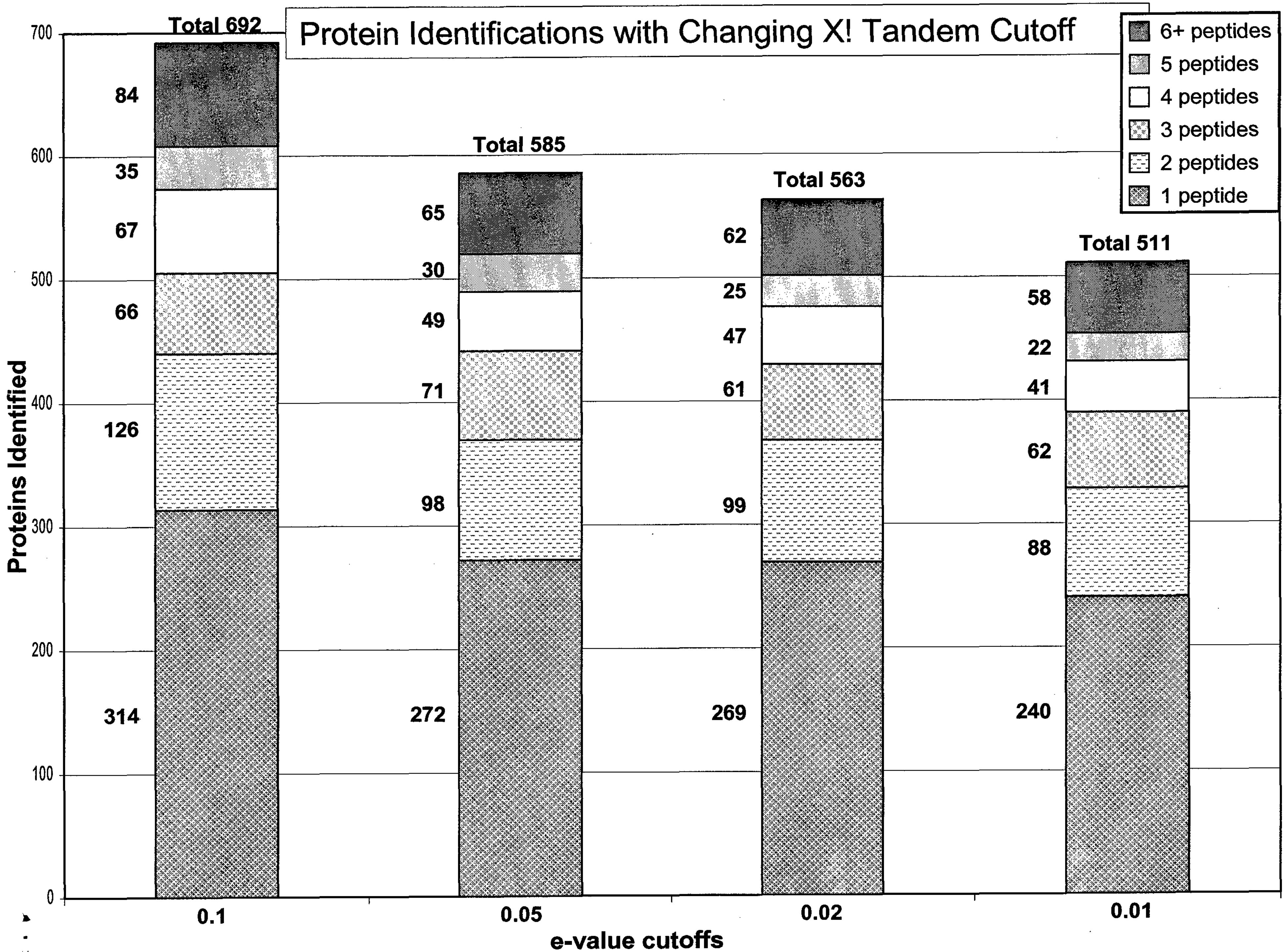

Comparison of False Positive Rates Between SEQUEST Cutoff Scores
3+ peptides
2 peptides
1 peptide
total
False Positive Rate
100%
95%
90%
85%
80%
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
1
2
3
4
5
Reference

Figure 3. Application of the X! Tandem search algorithm on yeast MudPIT data to determine an appropriate cutoff score (expectation value) for peptide matches to correlate with the output from a SEQUEST search, using our standard (Ref 5) parameters.

Fig 3



Protein Identifications with Changing X! Tandem Cutoff
6+ peptides
5 peptides
4 peptides
3 peptides
2 peptides
1 peptide
Proteins Identified
0
100
200
300
400
500
600
700
Total 692
84
35
67
66
126
314
Total 585
65
30
49
71
98
272
Total 563
62
25
47
61
99
269
Total 511
58
22
41
62
88
240
0.1
0.05
0.02
0.01
e-value cutoffs

OCT 21 2005

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	20.Oct.05	MAJOR REPORT

4. TITLE AND SUBTITLE
VERIFICATION OF SINGLE-PEPTIDE PROTEIN IDENTIFICATIONS BY THE APPLICATION OF COMPLEMENTARY DATABASE SEARCH ALGORITHMS.

5. FUNDING NUMBERS

6. AUTHOR(S)
MAJ ROHRBOUGH JAMES G JR

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF ARIZONA

8. PERFORMING ORGANIZATION REPORT NUMBER
CI04-1688

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
THE DEPARTMENT OF THE AIR FORCE
AFIT/CIA, BLDG 125
2950 P STREET
WPAFB OH 45433

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Unlimited distribution
In Accordance With AFI 35-205/AFIT Sup 1

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

14. SUBJECT TERMS	15. NUMBER OF PAGES
	28
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT